UDC 631.417.1:631.445.4(292.485:477.5)

Pedotransfer functions for prediction of soil organic carbon content for Chernozems Haplic and Calcic: A Case Study from the Left-Bank Forest-Steppe of Ukraine

Y. M. Dmytruk[1,2 a]✉, R. P. Palamarchuk[1 b], Y. F. Zhukova[1 c], N. V. Stepanenko[1 d]

[1] **State Institution "Soil Protection Institute of Ukraine",** *Kyiv, Ukraine*;
[2] **HEI "Podillya State University"**, *Kamianets-Podilskyi, Ukraine*

ARTICLE INFO

Received 12.05.2025
Received in revised form 14.06.2025
Accepted 17.06.2025
Available online 01.07.2025

Keywords:

soil organic Carbon; Chernozems Haplic and Calcic; pedotransfer function

ABSTRACT

The aim of the present research is to simulate and evaluate pedotransfer functions (PTF) for predicting soil organic carbon in Chernozem Haplic and Chernozem Calcic of the Left-Bank Forest-Steppe of Ukraine using data from agrochemical soil certification. Soil samples were collected from the upper arable horizon (0–20 cm) within the Left Bank Upland Province of the Forest-Steppe zone for Chernozem Haplic and Chernozem Calcic. According to the Ukrainian national soil classification system, the studied soils are classified as Typical Chernozems and Ordinary Chernozems. In addition to standard agrochemical analyses made by Poltava branch of State Institution «Institute for Soil Protection of Ukraine», in the Kyiv laboratory of this Institution determined the total carbon (TC), total inorganic carbon (TIC), and total organic carbon (TOC) contents. The Carbon content was determined using an analyzer Multi EA 4000 analyzer (Jena) via combustion. The total carbon released was automatically measured using a non-dispersive infrared detector (NDIR). The TOC content was determined as: TOC = TC - TIC (g/kg of soil). The descriptive statistical analysis and correlation and regression analyses were carried out using Excel; to assess the accuracy of the models and the agreement between the predicted and measured values of TOC (RMSE, NRMSE, MAE, R^2 and MAD). The statistically significant of the differences between the groups of predicted values of the TOC content was assessed using the t-test: Two-Sample Assuming Unequal Variances. The Humus and TOC content practically follow a normal distribution, which allows us to correctly estimate confidence intervals and significance levels; apply parametric statistical tests. The Humus content did not appear to be a key factor driving the TOC content. Taking into account the wide variability in TOC-to-Humus ratios observed in our study (ranging from 0.495 to 0.92), the use of a conventional conversion factor (e.g., 0.58) to estimate TOC content from Humus content may lead to substantial inaccuracies. The predicted values of TOC content using the PTFs, can be considered statistically interchangeable with the observed ones under the tested conditions, thereby confirming the adequacy of the models for practical application in TOC estimation. Given the absence of bias and the acceptable levels of prediction error (NRMSE < 10% for all models), the tested PTFs may be reliably used for TOC assessment in monitoring, land-use planning, and carbon stock estimation. Our vision of the practical use of modeled PTFs is as follow: if no exact information on the taxonomic classification of soils is available, it is better making use of PTF3 (TOC (g/kg) = 11.0+2.394*Humus - 0.049*P + 0.065*K); if the soil type is determined clearly, then for Chernozems Haplic – PTF5 (TOC (g/kg) = 32.1 + 1.21*Humus + 0.0017N + 0.088K - 0.085P - 2.71*pH), and for Chernozems Calcic – PTF7 (TOC (g/kg) = 15.4 + 1.81*Humus).

✉ dmytruk.yur@gmail.com

ORCID: [a]0000-0002-3157-0503, [b]0000-0002-5965-1305, [c]0000-0002-2755-5431, [d]0009-0005-6641-5887

Citing: Dmytruk, Y. M., Palamarchuk, R. P., Zhukova, Y. F., & Stepanenko, N. V. (2025). Pedotransfer functions for prediction of soil organic carbon content for Chernozems Haplic and Calcic: A Case Study from the Left-Bank Forest-Steppe of Ukraine. *AgroChemistry and Soil Science,* 98, 19-35. doi: 10.31073/acss98-02

1. Introduction

The global imperative to address climate change and promote sustainable agriculture has underscored the critical role of soil management practices in enhancing carbon sequestration and reducing greenhouse gas emissions. Statistical analyses and modeling are employed to quantify carbon sequestration rates, evaluate treatment effects, and assess environmental and agronomic benefits. Data collection and analysis techniques involve comprehensive soil sampling, greenhouse gas flux measurements, soil health assessments, and crop yield monitoring over multiple growing seasons. Statistical analyses and modeling approaches are employed to quantify the carbon sequestration rates, assess treatment effects on soil health and greenhouse gas emissions, and evaluate the overall environmental and agronomic benefits of the soil management practices [1].

Soil organic Carbon (SOC) is as a key determinant of yield in Europe. But the complex SOC-yield relationship highlights the necessity for tailored soil management strategies that consider specific site conditions to optimize C storage and crop yield [2]. SOC plays a crucial role in soil fertility, crop productivity, and agricultural sustainability. Long-term application of organic manure significantly increases SOC content, which is positively correlated with crop yields up to a critical level of about 9 g kg^{-1} [3]. Higher SOC content is associated with improved soil gas transport capability, potentially enhancing root growth and carbon input [4]. Optimal levels of decomposable organic carbon range from 0.2% to 0.6%, contributing to high crop yields when combined with mineral fertilization [5]. SOC provides nutrients to plants and improves water availability, enhancing soil fertility and food productivity [6]. These findings highlight the importance of maintaining appropriate SOC levels through balanced organic and mineral fertilization to optimize crop productivity and environmental sustainability in arable lands. The capacity of soils to store carbon (C) has substantial implications as for climate change mitigation as sustainable global crop production

While dry combustion is considered the standard method, in situ analytical techniques are being developed to offer increased accuracy, precision, and cost-effectiveness [7]. A recent photocatalytic approach using a PeCOD analyzer combined with GIS technology has shown promising results for on-site SOC determination. This method revealed that finer soil materials have higher SOC content, and SOC decreases with soil depth. Land management practices, precipitation, and temperature significantly affect SOC formation and residence time. GIS mapping of SOC distribution can identify hotspot areas, informing sustainable farming practices and climate change mitigation strategies [8].

Raising the organic matter content of soils is desirable for guaranteeing that soils will be able to keep fulfilling essential functions and services in spite of fast-changing environmental conditions. There is a thought, since at least 90% of plant residues added to soils to increase their carbon content over the long term are mineralized relatively rapidly and are released as CO2 to the atmosphere, farmers would have to apply to their fields 10 times more organic carbon annually than what they would eventually expect to sequester [9]. Recent studies demonstrate that organic farming practices significantly increase SOC compared to conventional methods. A meta-analysis revealed higher SOC concentrations and mass storage under organic agriculture, with fine soil textures showing the greatest improvements [10]. Similarly, research in the Netherlands found that organic arable systems had higher SOC content, mineralization rates, and water-stable aggregation than conventional systems, though still lower than permanent pasture [11]. While organic farming generally improves soil biochemical properties, physical aspects like susceptibility to compaction should also be considered [11, 12]. These findings highlight the potential of organic farming practices to enhance soil carbon sequestration and contribute to climate change mitigation and sustainable agriculture.

To advance sustainable and resilient agricultural management policies, especially during land use changes, it is imperative to monitor, report, and verify SOC content rigorously to inform its stock. However, conventional methods often entail challenging, time-consuming, and costly direct soil measurements [13]. The availability of quality data determines the accuracy of the model; therefore, it is necessary to build a general dataset that includes a carefully selected range of variables that, as a whole, influence the values of the variable we predict [14].

Pedotransfer functions (PTFs) are mathematical models used to predict soil properties, including SOC content. Studies have shown that PTFs can effectively predict SOC in subsurface horizons across Europe [15]. Machine learning techniques have been employed to develop generalized PTFs for predicting soil electrical conductivity and SOC content in Australia [16]. However, the choice of PTF for bulk density can significantly impact SOC storage estimates at regional scales, with uncertainties ranging from 10.61 % to 70.46 % in Chinese terrestrial ecosystems [17]. Key predictors for SOC content include soil depth, pH, soil texture, and geomorphology [16]. While PTFs show promise for estimating SOC content across various landscapes, careful consideration of the selected PTF and its limitations is crucial for accurate regional-scale assessments.

PTFs were developed to predict organic carbon content in subsurface horizons of European soils. The PTFs showed high reliability with adjusted R2 values above 0.62 and good predictive ability with RMSE values between 2.43 and 13.82 g C kg^{-1} soil. The use of bulk density as a predictor variable improved the performance of the PTFs, and they are suitable for continental-scale applications in Europe. Employed multiple linear regression with predictor variables: SOC in topsoil, depth of subsurface horizons, texture, bulk density, mean annual temperature, and precipitation [15].

PTF model types included: Regression-based models; Machine learning models; Process-based models and Additive/mixed statistical models. For forming PTF Models measured different soil properties: SOC content and SOC stock; SOC fractions (including humic substances, stable SOC). Other properties included SOC dynamics, total carbon and total nitrogen, bulk density, SOM and labile carbon. Various indicators were used to assess the effectiveness of the models, such as: R^2 (or adjusted R^2) and RMSE. The coefficient of determination (R^2) was used to assess the goodness of fit between predicted and measured values. Higher R^2 values indicated better model performance. Other metrics included RPD, MAE, MSE, ME, Willmott's d, and normalized RMSE [13, 18]. We should note that in recent studies, the mean absolute error (MAE), root mean square error (RMSE), and their normalized versions are widely used to validate predictive models in soil science research. These metrics are commonly used to evaluate model performance, and RMSE is more sensitive to outliers and potentially more suitable for Gaussian error distributions [19, 20]. Recently, researchers have applied these metrics to evaluate machine learning algorithms for digital soil mapping, achieving high accuracy in predicting soil nitrogen content [21]. In addition, these metrics were used to compare the effectiveness of different statistical approaches in predicting soil properties based on mid-infrared spectral data [22]. Note that in some cases, RMSE can be misleading in understanding the uncertainty of the forecast [19, 22]. Some authors [23] argue for RMSE as the optimal metric for normal errors, refuting the idea that MAE should be used exclusively. They do not contend RMSE is inherently superior and instead advocate that a combination of metrics, including both RMSE and MAE, should be used to evaluate model performance.

So, key predictors that are most often used in the researches analyzed by us are [18, 24, 25]: Soil properties (such as pH, texture, carbon/nitrogen ratio, cation exchange capacity, nitrogen, phosphorus, fertility); Spectral predictors (hyperspectral or MIR spectra); Environmental factors: topographic variables (curvature, elevation, aspect, wetness, valley depth, catchment area); Land use or management variables (ploughing, land use, reclamation); Climate variables (temperature, ESPI, nitrogen deposition). The authors recognized that the limiting factors may include related to predictor availability, sample size, model suitability for nonlinear data, and calibration dependencies [26, 27]. Mitigation strategies of the forecasting results provide for: data expansion (auxiliary predictors, open data, satellite/digital elevation model); model selection or optimization (nonlinear models, parameter optimization); laboratory or sample standardization; management recommendations [13, 28].

Determining the spatio-temporal dynamics of SOC content and stocks remains a critical task for ensuring food security, addressing climate change challenges, and halting land degradation [29, 30], which aligns with the United Nations Sustainable Development Goals [31]. Achieving these objectives is unattainable without systematic soil monitoring, which is currently lacking in Ukraine. In particular, there is no operational national system for observing soil properties over time and space, including SOC – one of the key indicators of soil fertility and ecosystem resilience [32]. As a result, the only available source of soil-related data for decision-making at all levels of agro-environmental governance in Ukraine is the agrochemical certification of farmlands. However, this certification does not include measurements of SOC, and given the ongoing war and limited government funding, a fundamental reform of the SOC monitoring system is not anticipated in the near future.

To overcome the information gaps on the soil organic carbon content, the modeling based on available agrochemical indicators is of crucial importance. A database of analytical data on SOC content for specific locations and for soils of different genesis should become a platform for modeling.

The aim of the present research is to simulate and evaluate pedotransfer functions for predicting soil organic carbon content in Chernozems Typical and Ordinary (Chernozem Haplic and Chernozem Callic accordingly) of the Left-Bank Forest-Steppe of Ukraine using data from agrochemical soil certification.

Objectives of the study: 1. To analyze the total organic Carbon (TOC) content of and agrochemical parameters (Humus content, available Phosphorus, Potassium, alkaline hydrolyzed Nitrogen, Soil acidity) in soil samples collected from the arable horizon of the study area. 2. Using both individual and complexes of soil indicators as predictors, build a series of PTF for predicting the content of TOC. 3. For different variants, to give a comparative assessment of the models' effectiveness in terms of Predictive power and Explanatory capability using statistical criteria (MAE, RMSE, NRMSE, R^2 and MAD). 4. To identify the possibilities of applying the constructed PTFs for practical use in the context of supporting decisions on sustainable soil resource management.

Author's contribution: Y. Dmytruk – idea, selection of methods, writing, conceptualization; Y. Zhukova – calculations, description of results, methodology; R. Palamarchuk – editing, synthesis and conclusions; N. Stepanenko – analytical research, software.

2. Materials and Methods

For agrochemical certification, soil samples were collected from the upper arable horizon (0–20 cm) within the Left Bank Upland Province of the Forest-Steppe zone (according to the agro-soil zoning of Ukraine). This territory corresponds to the Eastern Poltava Upland physical-geographical region and is classified as a zone of insufficient moisture in agro-climatic zoning (Hydrothermal Coefficient = 1,0–1,3). According to the Ukrainian national soil classification system, the studied soils are classified as Typical Chernozems and Ordinary Chernozems. These subtypes conditionally correspond to Chernozems Haplic and Chernozems Calcic, respectively, within the WRB framework (IUSS Working Group WRB, 2022), based on the presence or absence of a diagnostic calcic horizon and associated morphological and chemical properties.

In addition to standard agrochemical analyses [33] including Humus content, hydrolyzable Nitrogen (N), available Phosphorus (P), available Potassium (K), and Actual acidity (pH in water) made by Poltava branch of State Institution «Institute for Soil Protection of Ukraine»; in the Kyiv laboratory of the State Institution «Institute for Soil Protection of Ukraine» determined the total carbon (TC), total inorganic carbon (TIC), and total organic carbon (TOC) contents in the same samples (Table 1). The dataset (n = 29) comprised two subtypes of Chernozems: Chernozem Haplic (n = 18) and Chernozem Calcic (n = 11).

The carbon content of soils was determined using an analyzer Multi EA 4000 analyzer (Jena) via combustion, during which carbon is oxidized to CO_2 in an oxygen-rich gas stream at temperatures above 900°C. TC was determined via combustion at 1200°C in an oxygen atmosphere. The total carbon released was automatically measured using a non-dispersive infrared detector (NDIR).

The previously collected soil samples were dried indoors to an air-dry state without sunlight at a temperature not exceeding 400C. After drying, inclusions (coarse debris, organic residues (plant roots, leaves, seeds, etc.)) were removed manually from the soil samples, if any. Subsequently, the soils were sifted through sieves with a 2-mm mesh diameter, and after obtaining a representative sample (weighing 200–300 g), it was sifted again, but on sieves with a mesh diameter of 0.76 mm in accordance with the requirements for using the specified multielement analyzer. Prior to the analysis, the Multi EA® 4000 was calibrated using calcium carbonate $CaCO_3$ (12% C) of analytical strength. Organic carbon was determined by the combustion method, in which soil carbon is oxidized to carbon dioxide (CO_2) in an oxygen-containing gas stream at a temperature above 900°C. To determine total carbon, a 300 mg sample was taken in duplicate. The combustion in an oxygen atmosphere was performed at a temperature of 1200°C. The amount of total carbon (TC) released was determined automatically using a non-dispersive infrared (NDIR) detector. The inorganic carbon (TIC) determination was carried out according to the requirements for the device using the appropriate module of the analyzer. TIC was determined by adding 1500 mg of a 40 % solution of orthophosphoric acid (H_3PO_4) to the soil sample. The carbon dioxide (CO2) released during the reaction was recorded by NDIR. The carbon content measurement in soil samples was controlled using CRM No. B2176 (Elemental Microanalysis) and CRM No. B2237 (Elemental Microanalysis) with a carbon content of 15.95 % and 2.75 %, respectively. The TOC content was determined as the difference between total carbon and inorganic carbon: TOC = TC - TIC. The results are expressed in g/kg of soil.

The statistical analysis and correlation and regression analyses were carried out using Excel: mean, standard deviation (SD), coefficient of variation (V, %), variance, Student's t-test ($p < 0.05$) were estimated; to assess the accuracy of the models and the agreement between the predicted and measured values of soil organic carbon the parameters used were: root mean square error (RMSE), normalized root mean square error (NRMSE), mean absolute error (MAE), coefficient of determination (R^2) and, as a complementary indicator, median absolute deviation (MAD). The use of MAE, RMSE, MAD, and R^2 simultaneously allows to cover different aspects of model quality: accuracy, robustness to outliers, and the explanatory power [19, 23]. The statistically significant of the differences between the groups of predicted values of the TOC content was assessed using the t-test: Two-Sample Assuming Unequal Variances. To evaluate the agreement between the predicted and measured values of soil TOC, both statistical accuracy metrics and hypothesis testing were employed. The predictive performance of the models was assessed using MAE, MAD, RMSE, NRMSE, and the coefficient of determination (R^2), while two-sample t-tests were used to determine whether differences between measured and predicted TOC values were statistically significant.

3. Results

3.1. Overall analysis of soil parameters

The values of soil indicators resulting from the analysis of the soil samples are presented in Table 1.

Table 1
Soil indicators of Chernozems

Samples №№	AISG*	TC**	TIC	TOC	Humus	N	P	K	pHaq
			*g/kg****		*%*		*mg/kg*		
P1	55e	22.7	2.22	20.48	3.70	147.2	115.5	105.5	6.10
P2	55e	22.17	1.31	20.86	3.72	135.1	105.0	103.0	6.0
P3	55e	22.705	0.88	21.825	3.16	114.8	81.0	118.3	6.20
P4	53д	20.29	0.21	20.08	3.51	132.4	122.0	110.0	6.10
P7	65e	24.445	1.53	22.915	3.62	98.2	136.6	149.8	6.20
P8	65e	25.06	1.60	23.46	4.16	130.2	145.2	96.6	6.20
P9	65e	20.935	0.79	20.145	2.61	87.5	62.5	80.5	6.0
P10	65e	20.795	0.74	20.055	3.30	107.4	95.7	86.3	6.10
P11	65e	20.835	0.79	20.045	3.69	145.1	82.1	77.8	6.10
P12	53д	19.945	0.84	19.105	3.30	112.8	81.5	78.5	6.0
P13	53д	20.035	0.88	19.155	3.12	124.5	66.0	52.0	6.20
P14	53д	17.7	0.027	17.673	2.81	97.8	63.0	70.3	7.10
P15	53д	17.75	0.021	17.729	3.31	118.4	112.0	90.7	6.10
P16	53д	17.415	0.023	17.392	3.51	110.4	125.7	94.3	6.20
P17	53г	17.42	0.022	17.398	2.97	127.8	121.0	96.3	6.10
P18	53г	17.735	0.023	17.712	2.95	125.1	116.0	95.0	6.0
P19	53г	17.505	0.024	17.481	3.18	110.2	70.5	53.0	6.0
P20	65д	21.705	0.026	21.679	3.57	110.0	83.5	79.6	6.0
P21	65д	21.635	0.026	21.609	3.06	136.2	98.0	122.3	6.0
P22	65д	22.095	0.024	22.071	3.22	119.2	96.2	96.6	5.70
P23	58e	22.13	0.026	22.104	3.62	117.2	96.3	100.7	5.80
P24	58e	21.52	0.012	21.508	3.25	115.1	78.0	103.7	5.80
P25	58e	19.62	0.016	19.604	2.51	91.4	81.0	98.0	6.0
P26	53д	20.93	0.041	20.889	2.95	95.2	94.7	70.3	5.80
P27	53д	20.81	0.03	20.78	2.60	88.9	191.0	194.0	5.80
P28	53д	21.08	0.021	21.059	2.29	75.4	67.7	68.0	5.90
P29	53д	16.905	0.021	16.884	2.59	78.5	90.3	71.0	6.0
P30	53д	15.965	0.019	15.946	2.79	95.7	100.0	78.3	5.90
P31	53д	16.205	0.021	16.184	2.56	92.3	93.7	69.3	6.0
Mean		20.2	0.42	19.79	3.16	112	99.0	93.4	6.0
Minimum		15.965	0.012	15.9	2.29	75.4	62.5	52.0	5.70
Maximum		25.06	2.22	23.5	4.16	147	191	194	7.10

* – agro-industrial soil groups; due to the fact that there is no AISG in English, AISG are shown in Ukrainian transcription; ** – Carbon content: total (TC), inorganic (TIC) and organic (TOC); *** – the units of measuring in the all next tables are the similar

The analysis of soil indicators with the use of descriptive statistics suggests that all Chernozems are soils with a considerable fertility potential.

The statistical parameters (Table 2) characterizing the distribution of key soil properties demonstrate varying degrees of dispersion and deviation from normality.

Nitrogen and Humus contents exhibit relatively low variability, as indicated by low standard deviations (19.2 and 0.45, respectively) and moderate coefficients of variation (Table 2). Both parameters display near-symmetric distributions, with skewness values close to zero (-0.03 for nitrogen and 0,02 for humus), and negative kurtosis values (-0.73 and -0.52, respectively), suggesting slightly platykurtic distributions. TOC is characterized by moderate variability (SD = 2.07), a near-symmetric distribution (Skewness = -0.25), and slightly platykurtic shape (Kurtosis = -0.96) statistically significant.

Table 2
Statistical parameters of the dataset (n = 29)

Parameters	*Nitrogen*	*Phosphorus*	*Potassium*	*Humus*	*pHaq*	*TOC*
Mean	112	99.0	93.4	3.16	6.05	19.79
Median	113	95.7	94.3	3.18	6.0	20.1
St. Deviation	19.2	28.1	28.5	0.45	0.24	2.07
Variance	368	789	812	0.20	0.06	4.30
Kurtosis	-0.73	2.78	4.77	-0.52	12.2	-0.96
Skewness	-0.03	1.32	1.70	0.02	2.81	-0.25
Minimum	75.4	62.5	52.0	2.29	5.70	15.9
Maximum	147.2	191	194	4.16	7.10	23.5

In contrast, Phosphorus and Potassium contents demonstrate higher variability, with Significant values of standard deviations and variances. Their positive Skewness values (1.32 and 1.70) and leptokurtic profiles (Kurtosis 2.78 and 4.77, respectively) indicate right-skewed distributions with a higher probability of extreme values (Table 2). Notably, the pHaq values exhibit a highly peaked and right-skewed distribution. This suggests the presence of outliers or concentration of values near the lower boundary.

The average Humus content across all soils was 3.16 ± 0,45 %, indicating a moderate level of variability. The distribution of values is close to normal (skewness ≈ 0.02; kurtosis ≈ –0.52), suggesting that the results are representative of the general population. According to the data obtained, Chernozems Calcic are characterized by a higher mean Humus content (3.33 ± 0.48%) and slightly greater spatial heterogeneity. The slight left-skewness (–0.27) indicates a prevalence of relatively high Humus values with a small number of lower values. In contrast, Chernozems Haplic are more homogeneous in terms of Humus content (mean 3.06 ± 0.41%), showing lower variability and a nearly symmetrical distribution (kewness ≈ –0,03).

Soil organic carbon is a key indicator of soil fertility, resistance to degradation, and organic matter accumulation. The average TOC content in all analyzed samples was 19.79 g/kg, with a standard deviation of 2.07, indicating moderate variability. The Skewness (–0.25) and Kurtosis (-0.96) suggest a slightly left-skewed and flattened distribution relative to the normal, which is typical for agroecosystems characterized by uneven organic fertilization regimes and crop diversity (Table 2).

In Chernozems Haplic, the mean TOC content (18.81 g/kg) was lower than the overall average, with higher variability (σ = 1.87). This may be associated with the effects of intensive tillage, leading to humus mineralization and partial loss of organic carbon [34]. Nonetheless, the skewness (0.096) is close to zero, indicating a near-normal distribution, which supports the applicability of regression and ANOVA methods. In contrast, Chernozems Calcic exhibited a higher mean TOC content (21.38 g/kg) and lower dispersion (σ = 1.27) compared to Chernozems Haplic, indicating more homogeneous impact drivers on the plough layer. The Skewness was near zero (0.066), and the negative Kurtosis (–1.09) points to a slightly platykurtic distribution. Overall, TOC distributions in all soils studied approximate a normal distribution, thereby justifying the use of parametric statistical approaches for further analysis (Table 2).

The some soil indicators deviate from a normal distribution, particularly Phosphorus, Potassium, and pH. At the same time, the Humus content and TOC content practically follow a normal distribution, which allows us to correctly estimate confidence intervals and significance levels; apply parametric statistical tests (in particular, t-test, F-test). Additionally and important for us, the distribution's normality is a pre-requisite for using linear regression.

If the input data and residuals do not have significant deviations from normality, the model will be more stable and less sensitive to outliers, and the confidence intervals will be more precise.

To justify the selection of predictors for constructing PTFs aimed at estimating TOC content, pairwise correlation coefficients among soil properties were analyzed (Table 3). Correlation analysis provides an initial assessment of the strength and direction of linear associations between potential predictors and the target variable TOC, which is essential for informed model development. A moderate positive correlation was observed between TOC and both humus content ($r = 0.46$) and available Potassium ($r = 0.45$), suggesting their relevance as explanatory variables in the predictive models. In contrast, Nitrogen ($r = 0.23$), Phosphorus ($r = 0.16$), and soil pH ($r = -0.22$) showed weak or negligible correlations with TOC, indicating a limited linear predictive capacity in isolation (Table 3). Nevertheless, variables with low pairwise correlation should not be automatically excluded, especially in multifactor models, where interactions or nonlinear contributions may enhance model performance.

Table 3
Pairwise correlation coefficients (n = 29)

Parameters	N	P	K	Humus	pHaq	TOC
N	1.0					
P	0.15	1.0				
K	0.08	**0.77**	1.0			
Humus	**0.74**	0.30	0.13	1.0		
pHaq	0.06	-0.16	-0.16	0.09	1.0	
TOC	0.23	0.16	0.45	0.46	-0.22	1.0

3.2. Pedotransfer functions (models) for prediction of TOC

For the purpose of the study, we chose the local area of Chornozems Haplic and Calcic to prevent the need to take into account variability in climate, relief, and vegetation indices. The predictive power of the models varies significantly for different types of land use, so we chose only arable land. However, it is quite difficult to separate out the difference in the impact of management practices on TOC content.

The selection of an appropriate pedotransfer function (PTF) for practical application should strike a balance between predictive accuracy, explanatory capability, and operational simplicity. Ideally, models should incorporate as few predictors as possible to facilitate efficient field or laboratory use, without compromising reliability. On the other hand, increasing the number of input variables inevitably raises the model's complexity, not only due to the need for more extensive soil analysis but also because of the greater potential for cumulative analytical errors during data collection.

Based on a review of relevant literature (see Introduction), the most informative and commonly used predictors for estimating TOC content include humus (organic matter) content, particle-size distribution (texture fractions), total nitrogen content, bulk density, and its probable other properties [13–18], as well as the use, if possible, of spectral libraries, which is considered to be a perspective trend [35, 36].These factors should be prioritized depending on the specific context, available data, and analytical capacities.

To forecast the content of TOC in soils using the results of laboratory analyzes and agrochemical certification data, 9 PTFs were created (Table 4), of which four functions were built for the whole collection (without dividing soils into subtypes). The constructed PTFs based on linear regression demonstrate varying levels of effectiveness in estimating TOC content across different subtypes of Chernozems. The models differ in terms of input variables, determination coefficients (R^2), F-statistics, and statistical significance (p-values).

Table 4
Regression parameters and statistical characteristics of linear pedotransfer functions for TOC estimation

Pedotransfer Functions	a_0	Coefficients (b)	R^2	F	*p-value*
Chernozems **(n = 29)**					
PTF1 *p-value*: Humus – 0.013	13.13	2.106*Humus	0.21	7.12	0.013
PTF2 *p-value*: Humus – 0.0016; K – 0.0002; P – 0.004; N – 0.166	11.4	3.4*Humus+0.066*K - 0.05*P - 0.031N	0.57	7.85	0.0003
PTF3 *p-value*: Humus – 0.0013; K – 0.0004; P – 0.006	11.0	2.394*Humus - 0.049*P + 0.065*K	0.53	9.40	0.00025
PTF4 *p-value*: Humus – 0.016; K – 0.019	11.2	1.97*Humus+0.026*K	0.36	7.40	0.003
Chernozems Haplic **(n = 18)**					
PTF5 *p-value*: Humus – 0.45; K – 0.001; P – 0.004; N – 0.96; pH – 0.06	32.1	1.21*Humus + 0.0017N + 0.088K - 0.085P - 2.71*pH	0.63	4.1	0.021
PTF6 *p-значення*: K – 0.003; P – 0.016	18.3	0.077*K - 0.06*P	0.45	6.17	0.011
Chernozems Calcic **(n = 11)**					
PTF7 *p-value*: Humus – 0.019	15.4	1.81*Humus	0.47	8.12	0.019
PTF8 *p-value*: Humus – 0.21; K – 0.45; P – 0.22; N – 0.55; pH – 0.09	34.1	1.32*Humus - 0.01N + 0.01K + 0.03P - 3.33pH	0.85	5.6	0.04
PTF9 *p-value*: Humus – 0.19; K – 0.30; P – 0.54	13.9	1.26*Humus + 0.02K + 0.013P	0.71	5.79	0.026
Comment: a_0 – intercept; b – regression coefficients					

For both Chernozems (n = 29), PTF1, which relied solely on Humus content, exhibited limited predictive power (R^2 = 0.21), indicating that single-variable models may be insufficient for reliable TOC estimation in soils of diverse agricultural land. Although PTF4 improved performance slightly by adding available K (R^2 = 0.36), it still lagged behind the more comprehensive models. Among the developed pedotransfer functions, PTF2 and PTF3 showed the highest explanatory performance for estimating TOC content, with coefficients of determination (R^2) of 0.57 and 0.53, respectively. These models incorporated multiple routinely measured agrochemical indicators, including humus, available potassium (K), and phosphorus (P), all of which were statistically significant predictors ($p < 0.01$). The inclusion of these variables enhanced the models' ability to capture TOC variability across Chernozems. That is, explanatory capability of PTF3 increased by more than 30 % in terms of the determination coefficient R^2, which meets the criteria for high model accuracy [37]. Considering both model accuracy and practical applicability, PTF2 can recommended for use in agricultural land monitoring, particularly where full agrochemical dates (including humus, K, and P) are available. For cases with restricted laboratory access or where nitrogen data are not consistently reliable, PTF3 offers a more parsimonious yet sufficiently accurate alternative.

In the case of Chernozems Haplic, model PTF5 includes five predictors, but the p-value for humus ($p = 0.45$) and nitrogen ($p = 0.96$) suggests weak its individual contributions, despite the relatively high R^2 of 0.63. Conversely, PTF6, based only on K and P, shows better individual significance for both variables ($p = 0.003$ and $p = 0.016$ respectively), although with a slightly lower R^2 of 0.45 (Table 4).

For Calcic Chernozems, PTF7, which relies solely on humus, achieves a decent R^2 of 0.47 ($p = 0.019$). Model PTF8, which integrates five variables, demonstrates the highest explanatory power among all models ($R^2 = 0.85$), but some predictors (e.g., humus and nutrients) lack strong statistical significance ($p > 0.20$). PTF9, a reduced model excluding pH and N, still provides high predictive ability ($R^2 = 0.71$) with improved variable significance.

Thus, the inclusion of multiple predictors generally increases the determination coefficient; however, not all variables contribute significantly. Models such as PTF3 (Chernozems Haplic) and PTF9 (Chernozems Calcic) strike a balance between parsimony and predictive strength. The effectiveness of each model is context-dependent, and overfitting remains a risk when including statistically insignificant variables.

4. Discussion

4.1. Soil indicators

Descriptive statistics of the soil properties provide insights into the variability of TOC content across the studied sample set. The mean TOC concentration was 19.79 g/kg, with a standard deviation of 2.07 g/kg, indicating moderate variability within the dataset. In contrast, Humus content showed limited dispersion ($\sigma = 0.45$ %; range: 2.29–4.16 %), suggesting a relatively uniform this indicator among others.

The narrow range of humus variation likely contributed to its lower statistical influence in linear regression models. Although Humus is a primary source of TOC and proven relationship between carbon and Humus, its stable levels across samples may have diminished its predictive power in statistical modeling. On the other hand, more pronounced variability in the concentrations of available potassium (range = 142 mg/kg; $\sigma = 28.5$) and phosphorus (range = 128.5 mg/kg; $\sigma = 28.1$) was associated with their greater significance as predictors in multifactor models.

Additionally, the high skewness (Sk = 2.81) and excess kurtosis (Kurt = 12.25) observed in soil pH values highlight notable heterogeneity in the soil acid–base balance, which can substantially affect organic matter mineralization and carbon sequestration dynamics. This underscores the importance of including broader fertility indicators (e.g., mobile nutrient forms, pH) in predictive models for SOC, particularly in agroecosystems where Humus content remain within a narrow amplitude.

Therefore, the findings suggest that predictive models incorporating multiple soil fertility parameters—especially Potassium, Phosphorus, and pH – are statistically more robust and practically applicable in environments with relatively stable Humus content.

The Humus content did not appear to be a key factor in driving the TOC content, as a number of studies have shown. The reasons for this may be: a) limited variability of Humus content in the samples. If, for example, in Chernozems Haplic, the Humus content ranges within the limits insufficient to form a strong statistical dependence, its contribution may be devalued compared to more variable indicators (e.g., Potassium or Phosphorus); b) the effect of multicollinearity, since Humus correlates with Nitrogen

(Table 3) and therefore loses its individual significance in the multivariate model; the specific features of the soil texture were not taken into account, but the effect of specific fractions on the content of organic matter is known [41, 42].

As a reminder, there is a mismatch between soil organic matter (SOM) and Humus. In the vast majority of studies, correlations have been established between SOM and SOC, not between humus and SOC. In fact, the conversion factor of 0.58 is based on the assumption that soil organic matter (not humus) contains about 58 % carbon on average. But this is a topic for a scientific discussion of a different character than our planned tasks.

Based on our results the average TOC/Humus ratio across all Chernozems samples was 0.635 (ranging from 0.495 to 0.92), indicating substantial variation in the proportion of TOC within the organic matter pool (Table 5). Similar values were observed for Chernozems Haplic (mean: 0.625) and Chernozems Calcic (mean: 0,650) subtypes. Considering of previous studies reporting average conversion ratios between 0.55 and 0.60 [e.g., 36, 37], and underscores the necessity of site-specific calibration in SOC modeling.

Table 5
Statistical parameters of the Humus (%) and TOC (g/kg) content

Parameters	Chernozems (n = 29)		Chernozems Haplic (n = 18)		Chernozems Calcic (n = 11)	
	Humus	TOC	Humus	TOC	Humus	TOC
Mean	3.16	19.79	3.06	18,8	3,33	21,4
St. Error	0.08	0.385	0.096	0,44	0,14	0,38
Median	3.18	20.08	3.045	18,4	3,30	21,6
St.Deviation	0.45	2.07	0.41	1,87	0,48	1,27
Variance	0.20	4.30	0.17	3,49	0,23	1,60
Kurtosis	-0.52	-0.96	-0.72	-1,46	0,007	-1,09
Skewness	0.02	-0.25	-0.03	0,10	-0,27	0,07
Minimum	2.29	15.946	2.29	15,946	2,51	19,604
Maximum	4.16	23.46	3.72	21,825	4,16	23,46
Average TOC/Humus Ratio (min-max)	0.635 (0.495 – 0.92)		0.625 (0.495 – 0.92)		0.65 (0.54 – 0.78)	

Taking into account the wide variability in TOC-to-humus ratios observed in our study (ranging from 0.495 to 0.92), the use of a conventional conversion factor (e.g., 0.58) to estimate TOC content from Humus may lead to substantial inaccuracies. Our results clearly illustrate that such a generalized approach does not reflect the actual diversity of soil organic matter composition. Therefore, this underlines the critical need for site-specific or regionally calibrated studies aimed at improving the accuracy of TOC estimations, particularly in the context of carbon stock modeling and climate-smart soil management.

Taken together, these findings underline the importance of considering soil subtype when evaluating organic matter pools. The differences in both central tendency and variability between Haplic and Calcic Chernozems have implications for carbon budgeting, nutrient management, and the selection of pedotransfer functions in region-specific models.

4.2. The assessment of predictive power and explanatory capability of the PTF

Model agreement and statistical significance. Across all model variants tested, the results of the two-sample t-tests showed no statistically significant differences between the predicted and measured TOC values at the 95 % confidence level ($p > 0.05$ for all comparisons; Table 6). For instance, model PTF1 yielded a mean TOC prediction nearly identical to the measured value (19.784 vs. 19.787 g/kg, $p = 0.994$), while models PTF3 and PTF5 showed similarly close alignment ($p = 0.997$ and 0.974, respectively). Even in comparisons between model pairs, such as PTF3 vs. PTF1 or PTF7 vs. PTF9, no significant differences were observed ($p = 0.996$ and 0.21, respectively), indicating internal consistency among the modeling approaches (Table 6).

Table 6
Results of two-sample t-test comparing predicted and measured TOC values

Model	Mean, g/kg	Variance	t Stat	P(T<=t) two-tail	t Critical two-tail
PTF1 / measured	19.784 / 19.787	0.90 / 4.30	0.007	0.994	2.02
PTF3 / measured	19.786 / 19.787	2.29 / 4.30	0.003	0.997	2.01
PTF3 / PTF1	19.786 /19.784	2.29 / 0.90	0.004	0.996	2.01
PTF5 / measured	18.83 / 18.81	2.21 / 3.49	0.032	0.974	2.04
PTF6 / measured	19.16 / 18.81	1.64 / 3.49	0.66	0.514	2.04
PTF5 / PTF6	18.83 / 19.16	2.21 / 1.64	-0.72	0.475	2.03
PTF7 / measured	21.86 / 21.38	0.76 / 1.60	-1.04	0.312	2.10
PTF9 / measured	21.32 / 21.38	1.02 / 1.60	0.12	0.909	2.09
PTF7 / PTF9	21.86 / 21.32	0.76 / 1.02	1.30	0.210	2.09

Comment: t Stat – computed value of the t-statistic; P(T ≤ t) two-tail – two-sample p-value; t Critical two-tail – critical t-value for a two-sample t-test at $\alpha = 0.05$

Model accuracy across soil groups. In terms of predictive power, performance varied depending on the soil subtypes. For the total dataset ($n = 29$), model PTF3 outperformed PTF1, showing a lower RMSE (1.40 vs. 1.81 g/kg), lower NRMSE (7.06 % vs. 9.16 %), and higher R^2 (0.54 vs. 0.21), suggesting improved model generalizability (Table 7).

Table 7
Statistical Characteristics of the Derived Pedotransfer Functions

Parameters	PTF1*, PTF5, PTF7	PTF3, PTF6, PTF9
Chernozems (n=29)		
MAE, g/kg	1.61	1.07
MAD, g/kg	0.78	1.22
RMSE, g/kg	1.81	1.40
NRMSE, %	9.16	7.06
R^2	0.21	0.54
Model quality	Moderate	Good
Usability	Moderate	High
Chernozems Haplic		
MAE, g/kg	0.87	1.22
MAD, g/kg	0.95	0.89
RMSE, g/kg	1.10	1.40
NRMSE, %	5.87	7.42
R^2	0.63	0.45
Model quality	Good	Moderate
Usability	High	Moderate
Chernozems Calcic		
MAE, g/kg	0.71	0.54
MAD, g/kg	0.53	0.34
RMSE, g/kg	0.97	0.65
NRMSE, %	4.53	3.04
R^2	0.47	0.71
Model quality	Good	High
Usability	High	Good

Comment: PTF1, PTF3 – only for all Chernozems; PTF5, PTF6 – only Chernozems Haplic; PTF7, PTF9 – only Chernozems Calcic

When models were tested within soil subtypes, PTF5 (applied to Chernozems Haplic) demonstrated better performance than PTF6, with lower MAE (0.87 vs. 1.22 g/kg), lower RMSE (1.10 vs. 1.40 g/kg), and higher R^2 (0.63 vs. 0.45). Similarly, for Chernozems Calcic, PTF9 showed superior accuracy compared to PTF7, achieving a notably lower RMSE (0.65 vs. 0.97 g/kg), NRMSE (3.04 % vs. 4.53 %), and a higher R^2 (0.71 vs. 0.47).

These differences, although not statistically significant in terms of mean values, as confirmed by the t-test, may be practically important when selecting models for operational TOC estimation in soil subtypes. The greater predictive power of PTF9 and PTF5 in their respective domains highlights the advantage of tailoring models to concrete soil with similar physicochemical properties.

4.3. General implications

The overall absence of significant bias between predicted and measured TOC values, combined with moderate-to-high R^2 values (especially for subgtypes-specific models), confirms the feasibility of using these models for local-level TOC estimation. The highest predictive accuracy was achieved in Chernozems Calcic (R^2 up to 0.71), suggesting that TOC distribution in these soils may follow more consistent patterns compared to the broader dataset. The obtained results a two-sample t-test was applied for each model variant indicate an overall absence of statistically significant differences at the 0.05 level between the predicted and measured SOC values for all models considered (Table 6).

In particular, for the models PTF1, PTF3, and PTF5, the predicted means closely matched the measured TOC values (e.g., 19.784 vs. 19.787 g/kg for PTF1), with p-values significantly exceeding 0,05 ($p = 0.994$, 0.997, and 0.974, respectively). Similarly, comparisons between the model pairs (PTF3 vs. PTF1, PTF5 vs. PTF6, and PTF7 vs. PTF9) also revealed no significant differences, suggesting model stability and internal consistency across tested approaches.

Even the model PTF6, which, despite exhibiting a slightly higher mean TOC prediction (19.16 g/kg) compared to the measured value (18.81 g/kg), did not yield a statistically significant difference ($p = 0.514$). This further supports the robustness of the model under the current dataset.

It's interesting, models PTF7 and PTF9, which had the highest predicted means among all models tested (21.86 and 21.32 g/kg, p-values of 0.312 and 0.909, respectively), also showed non-significant deviations from the measured mean (21,38 g/kg). The direct comparison between these two models (PTF7 vs. PTF9) indicating the absence of bias between model variants predicting higher TOC contents.

Taken together, these results confirm that none of the tested models exhibit systematic bias relative to the measured TOC data. Therefore, the predicted values can be considered statistically interchangeable with the observed ones under the tested conditions, thereby confirming the adequacy of the models for practical application in TOC estimation. Given the absence of bias and the acceptable levels of prediction error (NRMSE < 10 % for all models), the tested PTFs may be reliably used for TOC assessment in agricultural monitoring, land-use planning, and carbon stock estimation, particularly when tailored to specific soil types.

The decision to choose a PTF model for practical application should take into account the best predictive power and explanatory capability of the model, and contain as few predictors as possible, which will ensure the efficiency of the model. At the same time, a larger number of predictors will a priori have a higher complexity both because of the need to determine the appropriate soil indicators and because of the increased possibility of errors (more analyzes - more probability of errors). Reducing the number of predictors helps minimize the influence of variability and potential analytical uncertainty in different scale assessments.

Future research should focus on expanding the model database across diverse soil-climatic zones, incorporating additional predictor variables (e.g., texture, bulk density, pH), and validating models using independent datasets under field conditions. Future work should explore the generalizability of the most accurate models (e.g., PTF2, PTF9) to broader soil types and under varying climatic and topographic conditions. Additionally, expanding model calibration with remote sensing data (e.g., NDVI, LST) may improve estimation in data-scarce regions.

5. Conclusions

This study evaluated the predictive accuracy and statistical validity of pedotransfer functions (PTFs) for estimating soil TOC content across Chernozems Haplic and Calcic. The following key conclusions can be drawn:

1. The models were assessed in terms of their robustness, accuracy, and predictive ability across diverse soils and input variables. No statistically significant differences were observed between predicted and measured TOC values for any of the models, as confirmed by two-sample t-tests ($p > 0.05$). This indicates that the models provide unbiased estimates and are statistically consistent with observed TOC measurements.

2. Model performance varied across soil subtypes. Among models for all Chernozems, PTF3 demonstrated superior accuracy over PTF1 for the entire dataset. Soil-specific models yielded higher performance: PTF5 showed better accuracy for Chernozems Haplic than PTF6; PTF9 outperformed PTF7 in Chernozems Calcic, reaching the highest overall R^2 (0.71) and the lowest NRMSE (3.04 %). Models calibrated separately for Chernozems Haplic and Calcic showed notably better performance than models developed for the entire dataset. This highlights the importance of considering soil classification when developing or applying PTFs for TOC estimation.

3. The decision to choose a model for predicting TOC content is the prerogative of a researcher. Our vision of the practical use of modeled PTFs is as follow: if no exact information on the taxonomic classification of soils is available, PTF3 (balance of simplicity, Predictive ability and Power); for Chernozems Haplic, PTF5 (highest accuracy, takes into account pH, K, P indicators), and for Chernozems Calcic, PTF7 (Predictive performance and Predictive ability).

4. The proposed models have some restrictions: 1) our sample concerned two soil subtypes in a specific location, which may affect the external validity of the models; 2) the factor that narrows the representativeness of pedotransfer functions is the quality of input data, both in terms of the accuracy of laboratory measurements and the representativeness of field soil sampling. The results should be interpreted taking into account the existing limits of PTF applicability.

Therefore, for practical TOC prediction, we recommend using adapted PFTs for the respective soils, taking into account the importance of the inputs and the ease of application.

References

1. Kaur, M., Nagabhooshanam, N., & Sharma, P. (2024). Quantifying the carbon sequestration potential of different soil management practices aimed at increasing organic content in soil and reducing the usage of chemical inputs. *Global NEST Journal*, 26(7). https://doi.org/10.30955/gnj.05932

2. Campos-Cáliz, A., Valencia, E, Plaza, C., Garland, G., Edlinger, A., Herzog, C., … García-Palacios, P. (2024). The positive effects of soil organic carbon on European cereal yields level off at 1.4 %. *Journal of Sustainable Agriculture and Environment*, 3, e70017. https://doi.org/10.1002/sae2.70017

3. Xie, E., Chen, J., Peng, Y., Yan, G., & Zhao, Y. (2024). Historical and future dynamics of cropland soil organic carbon stocks in an intensive human-impacted area of southeastern China. *Agriculture, Ecosystems & Environment*, 372*,* 109098. https://doi.org/10.1016/j.agee.2024.109098

4. Colombi, T., Walder, F., Büchi, L., Sommer, M., Liu, K., Six, J., … Keller, T. (2019). On-farm study reveals positive relationship between gas transport capacity and organic carbon content in arable soil. *SOIL*, 5(1), 91–105. https://doi.org/10.5194/soil-5-91-2019

5. Körschens, M. (2002). Importance of soil organic matter (SOM) for biomass production and environment: A review. *Archives of Agronomy and Soil Science*, 48(2), 89–94. https://doi.org/10.1080/03650340214162

6. Kanaujia, A., Gupta, J., Banerjee, S., & Malik, S. (2022). Correlation of organic carbon content of soil on paddy productivity: A case study. *International Journal of Advances in Agricultural Science & Technology*, 9(3), 43–46. https://doi.org/10.47856/ijaast.2022.v09i03.003

7. Chatterjee, A. Lal, R., Wielopolski, L., Martin, M. Z. & Ebinger, M. H. (2009). Evaluation of different soil carbon determination methods. *Critical Reviews in Plant Sciences*, 28(3), 164–178. https://doi.org/10.1080/07352680902776556

8. Abu El Haija, K., Chiang, Y. W., & Dos Santos, R. M. (2024). On-site determination of soil organic carbon content: A photocatalytic approach. *Clean Technologies*, 6(2), 784–801. https://doi.org/10.3390/cleantechnol6020040

9. Berthelin, J., Laba, M., Lemaire, G., Powison, D., Tessier, D., Wander, M., & Baveye, P. C. (2022). Soil carbon sequestration for climate change mitigation: Mineralization kinetics of organic inputs as an overlooked limitation. *European Journal of Soil Science*, 73(1). https://doi.org/10.1111/ejss.13221

10. Zhao, T., Kubota, H., Hernandez-Ramirez, G. (2024). Contrasting soil organic carbon concentrations and mass storage between conventional farming and organic farming: A meta-analysis. *Sustainability*, 16(24), 11260. https://doi.org/10.3390/su162411260

11. Pulleman, M. M., Jongmans, A. G., Marinissen, J. C. Y., & Bouma, J. (2003). Effects of organic versus conventional arable farming on soil structure and organic matter dynamics in a marine loam in the Netherlands. *Soil Use and Management*, 19(2), 157-165. https://doi.org/10.1079/SUM2003183

12. Sheoran, S., Meena, R. S., & Sharma, D. K. (2019). Impact of organic and conventional farming practices on soil quality: A global review. *Applied Ecology and Environmental Research*, 17(1), 951-968. https://doi.org/10.15666/aeer/1701_951968

13. Segura, C., Neal, A. L., Castro-Sardiňa, L., Harris, P., Rivero, M. J., Cardenas, L.M., & Irisarri, J. G. N. (2024). Comparison of direct and indirect soil organic carbon prediction at farm field scale. *Journal of Environmental Management*, 365*,* 121573. https://doi.org/10.1016/j.jenvman.2024.121573

14. Wadoux, A. M. J.-C., Minasny, B., McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210*,* 103359. https://doi.org/10.1016/j.earscirev.2020.103359

15. Fernandez-Ugalde, O., Tóth, G. (2017). Pedotransfer functions for predicting organic carbon in subsurface horizons of European soils. *European Journal of Soil Science*, 68*,* 716–725. https://doi.org/10.1111/ejss.12464

16. Benke, K. K., Norng, S., Robinson, N. J., Chia, K., Rees, D. B., & Hopley, J. (2020). Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. *Geoderma*, 366*,* 114210. https://doi.org/10.1016/j.geoderma.2020.114210

17. Xu, L., He, N. P., Yu, G. R., Wen, D., Gao, Y., & He, H. L. (2015). Differences in pedotransfer functions of bulk density lead to high uncertainty in soil organic carbon estimation at regional scales: Evidence from Chinese terrestrial ecosystems. *Journal of Geophysical Research: Biogeosciences*, 120(8), 1567-1575. https://doi.org/10.1002/2015JG002929

18. Hateffard, F., Szatmári, G., & Novák, T. (2023). Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. *Soil Science Annual*, 74(1), 165879. https://doi.org/10.37501/soilsa/165879

19. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7*,* 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

20. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15*,* 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

21. Auzzas, A., Carpa, G. F., Jani, A. D., & Ganga, A. (2024). An improved digital soil mapping approach to predict total N by combining machine learning algorithms and open environmental data. *Modeling Earth Systems and Environment*, 10*,* 6519–6538. https://doi.org/10.1007/s40808-024-02127-8

22. Dangal, S. R. S., Sanderman, J., Wills, S., & Ramirez-Lopez, L. (2019). Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Systems*, *3*(1), 11. https://doi.org/10.3390/soilsystems3010011

23. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7*,* e623. https://doi.org/10.7717/peerj-cs.623

24. Voltr, V., Menšík, L., Hlisnikovský, L., Hruška, M., Pokorný, E., & Pospíšilová, L. (2021). The soil organic matter in connection with soil properties and soil inputs. *Agronomy*, 11*,* 779. https://doi.org/10.3390/agronomy11040779

25. Zhang, H. Wu, P., Yin, A., Yang, X., Zhang, M., & Gao, C. (2017). Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Science of the Total Environment*, *592*, 704–713. https://doi.org/10.1016/j.scitotenv.2017.02.146

26. Ribeiro, S. G., Oliveira, M., Lopes, L. M., Costa, M. C. G., Toma, R. S., Araújo, I. C. S., … Teixeira, A. S. (2023). Reflectance spectroscopy in the prediction of soil organic carbon associated with humic substances. *Revista Brasileira de Ciência do Solo*, 47*,* e0220143. https://doi.org/10.36783/18069657rbcs20220143

27. John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., & Marcus Afu, S. (2020). Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land*, 9*,* 487. https://doi.org/10.3390/land9120487

28. Ťupek, B., Ortiz, C. A., Hashimoto, S., Stendahl, J., Dahlgren, J., Karltun, E., & Lehtonen, A. (2016). Underestimation of Boreal Soil Carbon Stocks by Mathematical Soil Carbon Models Linked to Soil Nutrient Status. *Biogeosciences*, 13, 4439–4459. https://doi.org/10.5194/bg-13-4439-2016

29. Lal, R. (2008). Carbon sequestration. *Philosophical Transactions of the Royal Society: Biological Sciences,* 363(1492), 815-830. https://doi.org/10.1098/rstb.2007.2185

30. Lal, R. (2015). Restoring Soil Quality to Mitigate Soil Degradation. *Sustainability*, 7, 5875-5895. https://doi:10.3390/su7055875

31. Klaus, L., Lal, R., Ehlers, K. (2019). Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations' Sustainable Development Goals. *Land Degradation & Development,* 30(7), 824-838. https: //doi.org: 10.1002/ldr.3270

32. Romanova, S., Dmytruk, Y., Zhukova, Y. (2024). Soil monitoring infrastructure in response to war. *International Journal of Environmental Studies,* 81(1), 189–198. https://doi.org/10.1080/00207233.2024.2314892

33. Yatsuk, I. P., Balyuk, S. A. (Eds.). (2019). *Methodology for Agricultural Chemical Certification of Agricultural Land: Guidance Document.* Kyiv [In Ukrainian].

34. Haddaway, N. R., Hedlund, K., Jackson, L. E., Kätterer, T., Lugato, E., Thomsen, I. K., … Isberg, P.-E. (2017). How does tillage intensity affect soil organic carbon? A systematic review. *Environmental Evidence, 6:30.* https://doi.org/10.1186/s13750-017-0108-9

35. Nocita, M., Stevens, A., Toth, G., Panagos, P., Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry,* 68, 337-347. https://doi.org/10.1016/j.soilbio.2013.10.022

36. Grunwald, S., Congrong, Y.U., & Xiong, X. (2018). Transferability and Scalability of Soil Total Carbon Prediction Models in Florida, USA. *Pedosphere,* 28(6), 856-872. https://doi.org/10.1016/S1002-0160(18)60048-7

37. Willmott, C. J. & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research,* 30, 79-82. http://dx.doi.org/10.3354/cr030079

38. Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Chapter One – Digital Mapping of Soil Carbon. *Advances in Agronomy, 118, 1-47.* https://doi.org/10.1016/B978-0-12-405942-9.00001-3

39. Jensen, J. L., Christensen, B. T., Schjønning, P., Watts, C. W., & Munkholm, L. J. (2018). Converting loss-on-ignition to organic carbon content in arable topsoil: pitfalls and proposed procedure. *The European Journal of Soil Science,* 69(4), 604-612. https://doi.org/10.1111/ejss.12558

40. Prout, J. M., Shepherd, K. D., McGrath, S. P., Kirk, G. J. D., & Haefele, S. M. (2021). What is a good level of soil organic matter? An index based on organic carbon to clay ratio. *The European Journal of Soil Science,* 72(6), 2493-2503. https://doi.org/10.1111/ejss.13012

41. Six, J. Conant, R. T., Paul, E. A. & Paustian, K. (2002). Stabilization Mechanisms of Soil Organic Matter: Implications for C-Saturation of Soils. *Plant Soil,* 241, 155-176. https://doi.org/10.1023/A:1016125726789

42. Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., Lützow, M., Marin-Spiotta, E. … Kögel-Knabner, I. (2019) Soil organic carbon storage as a key function of soils – A review of drivers and indicators at various scales. *Geoderma,* 333, 149-162. https://doi.org/10.1016/j.geoderma.2018.07.026

УДК 631.417.1:631.445.4(292.485:477.5)

Педотрансферні функції для прогнозування вмісту органічного вуглецю в чорноземах типових і звичайних на прикладі Лівобережного Лісостепу України

Ю. М. Дмитрук[1,2] ✉, Р. П. Паламарчук[1], Я. Ф. Жукова[1], Н. В. Степаненко[1]

[1] **Державна установа «Інститут охорони ґрунтів України»,** *Київ, Україна*
[2] **ЗВО «Подільський державний університет»**, *Кам'янець-Подільський, Україна*

✉ dmytruk.yur@gmail.com

Отримано 12.05.2025; Доопрацьовано 14.06.2025; Затверджено 17.06.2025; Доступно online 01.07.2025

Розширена анотація

Наразі в Україні моніторинг ґрунтів у його науковому розумінні відсутній, а реальним контролем стану ґрунтів на землях с.-г. призначення є агрохімічна сертифікація. Проте її методика не передбачає вимірювання вмісту органічного вуглецю у ґрунті (ОВҐ) і, враховуючи триваючу війну та обмежене державне фінансування, фундаментальна реформа системи моніторингу ґрунтів найближчим часом не очікується. Для подолання інформаційних прогалин щодо кількості та запасів ОВҐ вирішальне значення має моделювання на основі доступних агрохімічних показників. Метою цього дослідження є створення з використанням даних агрохімічної паспортизації ґрунтів та оцінка педотрансферних функцій (ПТФ) для прогнозування вмісту ОВҐ у чорноземах типових і звичайних Лівобережного Лісостепу України. *Цілі дослідження*: 1. З використанням сучасного аналізатора Multi EA 4000 в зразках ґрунтів, відібраних з орного горизонту, визначити вміст ОВҐ та агрохімічні показники (вміст гумусу, доступних форм фосфору та калію, лужногідролізованого азоту, pH_{H_2O}); 2. Як предиктори для моделювання серії ПТФ в цілях прогнозу вмісту ОВҐ використати вказані ґрунтові показники як окремі, так і їх поєднання; 3. Для створених варіантів ПТФ провести порівняльне оцінювання ефективності моделей з огляду на їхню прогностичну силу та пояснювальну здатність, використовуючи статистичні критерії (MAE, RMSE, NRMSE, R^2 та MAD); 4. Визначити можливості застосування побудованих ПТФ для практичного використання в контексті підтримки рішень щодо сталого управління ґрунтовими ресурсами. *Матеріали та методи.* У процесі агрохімічної сертифікації зразки ґрунтів було відібрано з верхнього (орного) горизонту (0–20 см) у межах Лівобережної височинної провінції Лісостепової зони для двох підтипів чорноземів: типовий (n = 18) та звичайний (n = 11). Окрім стандартних агрохімічних аналізів, здійснених в лабораторії Полтавської філії Державної установи «Інститут охорони ґрунтів України», у київській лабораторії цієї ж установи визначали загальний вміст вуглецю (TC), загальний вміст неорганічного вуглецю (TIC) та загальний вміст органічного вуглецю (TOC) у тих самих зразках ґрунтів. Вміст вуглецю визначали за допомогою аналізатора Multi EA 4000 (Jena) шляхом спалювання, під час якого вуглець окиснюється до CO_2 за різної температури в кисневій атмосфері. Загальний обсяг вивільненого вуглецю вимірювали автоматично за допомогою недисперсійного інфрачервоного детектора (NDIR). Загальний вміст органічного вуглецю (TOC) визначали як різницю між загальним вмістом вуглецю і вмістом неорганічного вуглецю: TOC = TC - TIC (г/кг ґрунту). Статистичний, кореляційний та регресійний аналізи проводили з використанням програми Excel для оцінювання точності моделей та узгодженості між прогнозованими та виміряними значеннями TOC. Статистично значущість відмінностей між групами прогнозованих значень вмісту TOC оцінювали за допомогою двовибіркового t-критерію з припущенням нерівних дисперсій. В оцінці узгодженості між прогнозованими та виміряними значеннями TOC використовували як статистичні показники точності, так і перевірку гіпотез. *Результати та висновки*. Вміст гумусу та ОВҐ (TOC) практично відповідають нормальному розподілу, що дозволяє нам правильно оцінювати довірчі інтервали та рівні значущості, застосовуючи параметричні статистичні тести. Крім того, що особливо важливо, нормальність розподілу є передумовою для надійного використання лінійної регресії. Також для обґрунтування вибору предикторів для побудови ПТФ, спрямованих на оцінку вмісту ОВҐ (TOC), були проаналізовані парні коефіцієнти кореляції між властивостями ґрунтів. Вміст гумусу не завжди був ключовим фактором, що впливає на вміст органічного вуглецю. Можливими причинами цього можуть бути: а) обмежена амплітуда вмісту гумусу у відібраних нами зразках; б) вплив мультиколінеарності, оскільки гумус достовірно корелює, наприклад, з умістом азоту, а тому втрачає своє індивідуальне значення в багатовимірній моделі; в) відсутні дані про гранулометричний склад ґрунтів, окремі фракції якого можуть мати значний вплив на органічну речовину ґрунтів. Беручи до уваги широку мінливість співвідношень TOC і гумусу, що спостерігалися в нашому дослідженні (від 0,495 до 0,92), використання звичайного коефіцієнта перетворення (0,58) для оцінки вмісту TOC на основі вмісту гумусу може призвести до суттєвих неточностей. Наші результати чітко ілюструють, що такий узагальнений підхід не відображає фактичної різноманітності складу органічної речовини ґрунтів. Це підкреслює критичну потребу в локальних дослідженнях, з урахуванням специфіки та структури ґрунтового покриву в конкретних ареалах, спрямованих на підвищення точності оцінок TOC, особливо в контексті моделювання запасів вуглецю та програми кліматично розумного управління ґрунтами. Прогнозовані значення вмісту TOC з використанням створених нами ПТФ можна вважати статистично взаємозамінними зі спостережуваними в протестованих умовах, що підтверджує адекватність створених нами моделей для практичного застосування в оцінці вмісту TOC. Враховуючи відсутність зміщення та прийнятні рівні похибки прогнозування (NRMSE <10 % для всіх моделей), протестовані ПТФ можуть бути надійно використані для оцінювання TOC у програмах моніторингу орних ґрунтів з наступною оцінкою запасів TOC, у плануванні заходів вуглецевого землеробства та вуглецевої сертифікації, особливо якщо за використання ПТФ будуть враховані конкретні таксони ґрунтів. Рішення про вибір моделі для прогнозування вмісту TOC є прерогативою дослідника. Наше бачення практичного використання змодельованих ПТФ є таким: якщо бракує точної інформації про класифікаційну приналежність ґрунтів, доцільно використовувати створену нами ПТФ3: TOC (г/кг) = 11,0 + 2,394Гумус - 0,049P + 0,065K; якщо точно визначено тип ґрунту, то для чорноземів типових оптимальним варіантом буде ПТФ5: TOC (г/кг) = 32,1 + 1,21Гумус + 0,0017N + 0,088K - 0,085P - 2,71pH, а для чорноземів звичайних — ПТФ7: TOC (г/кг) = 15,4 + 1,81Гумус. Вказані пропозиції не виключають можливості використання всіх створених нами ПТФ. Запропоновані моделі мають певні обмеження: 1) наша вибірка стосувалася двох підтипів ґрунтів у конкретних умовах, що може вплинути на зовнішню валідність моделей; 2) чинником, що звужує репрезентативність ПТФ, є якість вхідних даних, як зі сторони точності лабораторних вимірювань, так і щодо репрезентативності ґрунтових зразків. Змодельовані результати необхідно інтерпретувати з урахуванням існуючих обмежень у застосуванні ПТФ.

Ключові слова: органічний вуглець ґрунту; чорноземи типові і звичайні; педотрансферні функції

Форма цитування: Dmytruk Y. M., Palamarchuk R. P., Zhukova Y. F., Stepanenko N. V. Pedotransfer functions for prediction of soil organic carbon content for Chernozems Haplic and Calcic: A Case Study from the Left-Bank Forest-Steppe of Ukraine. *Агрохімія і ґрунтознавство :* міжвідом. темат. наук. зб. / ННЦ "ІҐА ім. О. Н. Соколовського*"*. Харків, 2025. Вип. 98. С. 19–35. doi: 10.31073/acss98-02